Combining linguistic and knowledge-based engineering for information retrieval and information extraction

Paul E. van der Vet and Bas van Bakel
Vossius Laboratory for Content Engineering
CTIT, University of Twente
P.O. Box 217, 7500 AE Enschede, the Netherlands
Phone +31 53 489 3694, fax +31 53 489 3503
Email {vet, bakel}@cs.utwente.nl

Abstract

Controlled-term indexing (the method of choice for multimedia collections and still very popular for purely textual material), appears an expensive solution because it takes huge resources and manual indexing. It is not possible, however, to perform a well-founded asessment of various approaches to information retrieval. We discuss ways to improve controlled-term indexing and illustrate these by looking at the Condorcet project carried out at Twente by us and co-workers. We round off with a discussion that, we hope, will raise more questions than it answers.

Keywords: Knowledge-Based Systems, Language Technology, Information Retrieval, Multimedia, Information Extraction

1 Introduction

Among the papers of the French mathematician and social philosopher Condorcet (1743–1794) found after his death, there is a proposal to assign each and every piece of knowledge a unique code. The code would serve two purposes: it could be used to organise libraries, and by comparing the codes with what was known, lacunae in knowledge could be discovered. Organisation of libraries was becoming a pressing problem in Condorcet's time because the production of printed material was turning into a flood. (Still, the antique library of Alexandria with its 700,000 manuscripts must have been a nightmare for newcomers,too.) Condorcet's proposal basically defines what we now call *controlled terms*: tokens taken from a pre-defined list with a meaning that is fixed with respect to that of the other tokens in the list. Then, as now, these terms were assigned by hand. One of the advantages of controlled terms, already noted by Condorcet, is their independence of medium in which the information is expressed. Controlled-term indexing therefore is the method of choice for multimedia collections.

We will first contrast controlled-term and uncontrolled-term approaches to find the relative pros and cons of each. The pressing question is: which method is preferred, relative to the situation? As we will show, this question cannot be answered for lack of data. We then turn to improvements of current controlled-term indexing practices. We discuss one technique in more detail by way of an overview of an information retrieval project carried out at our group, named after Condorcet. We round off with a discussion rather than conclusions.

2 The mystery of the continued use of controlled terms

About a decade ago, the Dutch sociologist Laeyendecker wrote a book under the title "Does progress bring us any further?" (our translation) [19]. For information retrieval (IR), many IR experts say 'yes' but not every end-user believes them. Probabilistic approaches are claimed to be the cheap and satisfactory solution to *ad-hoc* IR problems but many end-users stick to controlled terms. A probabilistic approach produces a document representation that consists of uncontrolled terms (basically, stemmed words or regularised phrases from the text itself minus the so-called stopwords). Controlled terms, by contrast, are taken from pre-defined resources such as thesauri and classification systems and need not occur in the document to which they are assigned. Given the impressive investments needed to make

controlled-term systems work, one wonders why such systems are still around and whether they perform as well as their users expect. These questions seem simple but turn out to be very difficult to answer.

The two approaches can be contrasted as if they were rivals, which they are not. (Any sensible system designer will offer users both possibilities.)

Controlled terms have two advantages. First, indexers and users alike at least share a common resource: the store of controlled terms. This reduces uncertainty. Second, because controlled terms are assigned by hand, they are completely media-independent. Texts in different languages about tigers, photographs and videos of tigers, and audio files with tiger sounds all receive the same controlled term, say, `tiger`. Against this, manual work is error-prone. When the *Chemical Abstracts* thesaurus was converted from hardcopy into an electronic version, many errors were introduced [23]. In our own investigations we have inspected *Engineered Materials Abstracts* and *Excerpta Medica* and found indexing errors, although there were fewer errors than we (biased computer scientists) initially thought. Further, it is a nuisance that proper names are seldom, if ever, declared to be controlled terms. In a number of retrieval situations we will want to search for proper names. Finally, since the development and maintenance of the store of controlled terms and indexing the documents have to be done by hand, a huge investment is needed. For comparison, Chemical Abstracts Services has close to 1,000 employees.

The big advantage of uncontrolled terms is low costs of indexing. Preparation of document representations can be fully automated, and costly investments like those for term resources are wholly avoided. The major disadvantages are: inability to abstract from the media used, and ambiguity of tokens: natural-language words or pictorial elements.

As regards effectiveness, there is ample material on uncontrolled-term systems. For controlled terms, there is no empirical material of a comparable quality and breadth. We simply do not know how well controlled-term systems perform, so a comparison between the two kinds of system on effectiveness is impossible.

In a famous experiment [3, 2], Blair and Maron measured the effectiveness of a STAIRS system that used uncontrolled-term indexing. They found the measured effectiveness disappointing and certainly below the requirements imposed by the situation. Searching was hindered by the very many ways the same subject can be characterised in natural language and even by cross-document anaphora (like "the subject of your last letter"). Blair and Maron concluded that it is simply infeasible for users to predict what words, word combinations or phrases would occur in the documents they sought and would not occur in the documents they did not seek. They advocated the use of controlled terms to enhance effectiveness, although they have not conducted a follow-up investigation to substantiate the claim. We find their argument plausible, at least for the situation and corpus investigated. For us, it is among the reasons to pursue a controlled-term approach in our own research.

Later experiments at TREC [34, 28] show that the situation has improved with respect to the figures found by Blair and Maron, but not spectacularly so. At the last TRECs, results seem to have reached a plateau [28]. From this, one cannot conclude that uncontrolled-term systems perform in an unsatisfactory way. After all, not every IR situation is as demanding as that investigated by Blair and Maron. We estimate that uncontrolled-term systems are a good choice for quite a number of applications.

The main problem in comparing approaches is that there are no estimates, and *a fortiori* no reliable estimates, of the total cost-benefit balance of an IR session. Benefits include expenses avoided and direct gains. Costs include:

1. Costs of setting up the system (including indexing), depreciated over sessions.
2. Costs of use, which can be broken down into:
   (a) Hardware costs (processing time, memory usage, network usage).
   (b) Costs of searching (handling some requests may take hours or even days of query construction and interactive refinement).
   (c) Costs of sifting the set of retrieved documents.
   (d) Costs incurred by missing relevant documents.

Item 2(a) can be safely neglected relative to the other cost items. The familiar measures of precision and recall bear on items 2(c) and 2(d) only and both the measurement and the subsequent interpretaion of these quantities is fraught with difficulties (see, for instance, [16]). Results in

terms of precision and recall at best represent an incomplete picture.

We simply do not know key items such as the costs of searching and costs of missing relevant documents. Just to illustrate how reliable cost-benefit figures, if they were available, would affect our judgments, consider a fictitious comparison between a controlled-term system and an uncontrolled-term system with identical recall and precision. The huge investments needed to get the controlled-term system into the air are earned back if the average session lasts significantly shorter than the average session on the uncontrolled-term system.

Cooper [7] has proposed to measure retrieval effectiveness in terms of the amount of money a person is willing to pay for having a system process an information request. Obviously, such a person has little to go on.

3 New approaches to controlled-term indexing

In our own research, we further explore controlled-term indexing. Computers can be employed in this approach, too, to obtain a more effective and efficient way of working. We investigate two improvements: better term resources and lowering of costs.

Term resources can be improved because current thesauri and classification systems are not very expressive. This state of affairs is due to the fact that, until recently, these resources had to be distributed and consulted in printed form. Computer manipulation opens new possibilities. A tangled hierarchy spanned by a number of different relations, for instance, becomes unreadable in printed form but is easy to understand and use with the help of computer programs. Modern jargon calls the computer-age successors of thesauri and classification systems *ontologies* [12, 22]. See [13, 32] for examples of ontologies that are too complex to be handled by other than automated means.

Ontologies allow indexers to assign *co-ordinated index terms* to documents to enable more precise searching. For example, suppose `aspirin` and `headache` are both controlled terms. With the help of co-ordination, we can specify the nature of the relation between these two terms in cases where they are both assigned to a document: for instance, `cures(aspirin, headache)` or `causes(aspirin, headache)`. Searchers can thus limit their search by specifying the relation. The query engine we developed for these terms [33] also handles generalisations, *e.g.*, `cures(any(medicine), headache)` will retrieve all documents about medicines against headache.

Costs can be lowered by partially automating the process of assigning controlled terms. (We say 'partially' because fully automatic assignment will not be both technically and economically feasible for a long time to come.) Text understanding and figure understanding are the fields that will have to spawn the necessary techniques. Figure understanding is a long way off, but text understanding is within reach. Indexing texts using text-understanding techniques is the subject of the next section. Documents in other forms will still have to be indexed by hand. The advantage of complete media-independence of controlled terms is not abandoned.

What remains are the costs of maintaining the resources. Resources are nowadays often simply lacking, so on the short term there are additional costs for making those resources in the first place. The resources include ontologies, grammars of natural languages, lexica that map natural-language words and phrases onto conceptual equivalents, knowledge bases with domain knowledge, and programs. We estimate that an ontology alone is more expensive than a thesaurus or classification system, so on the face of it this route only augments already substantial costs.

There are grounds, however, to think that many of the required resources will come into existence anyway. Unlike thesauri and classification systems, the resources required for semi-automatic indexing are also valuable for other applications. It is not difficult to foresee a future in which manipulation of information on the level of its content is commonplace. Workers in medicine have realised this earlier than their colleagues in other disciplines. The *Unified Medical Language System* (UMLS) [21] will grow into a body of resources that covers most of the needs of a semi-automated indexing system. Other disciplines will undoubtedly follow.

In our Condorcet project, we use UMLS as a resource. Basically, UMLS is a collection of thesauri, a lexicon that maps nouns and some phrases onto thesaurus terms, and a semantic network. The semantic network defines what are called *types* in a taxonomic hierarchy. Every term in every thesaurus is assigned a type to disambiguate meanings, for example, `cold` as indication of temperature *versus* `cold` as a disease.

From the thesauri, we use the MeSH Main Headings and the MeSH NM file (terms for chemicals). The combination of MeSH term and UMLS type is a concept in the sense of an ontology. The semantic network further defines about fifty relations that may hold between terms, depending on their types. We use these relations as co-ordinators to construct co-ordinated index terms such as `affects(zonisamide, epilepsy)` to index a document that discusses the use of zonisamide as anti-epileptic.

4 Condorcet

4.1 Overview

Condorcet (funded by the Dutch Technology Foundation (STW) through the *Werkgemeenschap Informatiewetenschap*, the Dutch Society for Information Science) focuses on semi-automatic indexing using controlled terms. We present an overview here; readers are referred to the Condorcet web site at

`http://www.cs.utwente.nl/condorcet/`

for more information and publications, including the three *Annual Reports* that have appeared so far.

Condorcet aims to build a prototype indexing system for large volumes of documents covering two scientific domains: *mechanical properties of engineering ceramics* as a field of materials science, and *epilepsy* as a subfield of medicine. Two domains rather than only one were chosen to avoid bias in the design of the indexing system. Ideally, when switching to another domain only the domain resource has to be changed. The documents in the development corpus are taken from machine-readable one year volumes of two bibliographic journals: the 1988 volume of *Excerpta Medica* from Elsevier Science Publishers, and the 1990 volume of *Engineered Materials Abstracts* from Materials Information. The prototype will be tested on 400 documents. Figures 1 and 2 present examples of document descriptions taken from the two sources. In the course of designing the system, we continuously incorporate techniques that enable the system to process much larger volumes, up to several hundred thousand documents.

Basically, indexing by Condorcet consists of mapping title plus abstract onto terms and co-ordinators by making intensive use of three kinds

AN: 88100203

TI: Effects of zonisamide in children with epilepsy

AB: The effects of zonisamide (1,2-benzisoxazole-3-methanesulfonamide: AD-810) were studied in 50 children with epilepsy, ranging in age from 3 months to 20 years (mean, 10.5 years). The types of epilepsy were primary generalized in one case, secondary generalized in 32, and partial in 17. The initial dose was 1-6 mg/kg/day and the dose was increased to 1.5-15 mg/kg/day. Four cases (8%) showed a complete disappearance of seizures and thirteen patients (26%) had a disappearance rate of 50% or more of seizures. Disappearance or improvement of seizures was obtained in 31% of the cases of generalized epilepsy and in 41% of the cases of partial epilepsy. Zonisamide was effective in 39% of cases of Lennox-Gastaut syndrome. Seizures completely disappeared in three of the four new cases. Spike discharges disappeared or significantly decreased in 22% of the cases that had undergone electroencephalograms. The blood levels of zonisamide were 10.8-18.8 mu g/ml in the three new cases when the seizures were controlled. Side effects such as drowsiness, ataxia, and salivation were observed in 42% of the children, more particularly in children receiving polypharmacy.

Figure 1: Part of a document description from the epilepsy domain, © Elsevier Science. 'AN' identifies the primary key, 'TI' and 'AB' the title and abstract parts. The present text reproduces the ASCII text as it is found in the file, hence, for instance, the string 'mu g/ml' for the SI unit μg/ml.

of knowledge. Knowledge of language and knowledge of the domain are combined to generate conceptual representations for the sentences in the document description, and indexing knowledge is used to generate index concepts from these conceptual representations. This indexing strategy is based on the idea of efficient use of the different kinds of knowledge. It is fully tuned to the objective of controlled-term indexing rather than focused on either linguistic or knowledge-based engineering, as is done in quite a number of other research projects [1, 11]. We return to this point below, in the discussion.

The problems involved in mapping document descriptions onto index terms and co-ordinators are linguistic problems and problems that involve inferences using domain knowledge. Therefore, combining linguistic and knowledge-based engineering appears a logical (but far from trivial) answer. Apart from how to make the combination conceptually, a more practical problem Condorcet tackles is how to design and develop a prototype indexing system that meets the design and

01 9001C1-C-0019
02 Influence of Ambient Temperature Sliding Velocity Under Unlubricated Sliding Conditions on Friction and Wear of Si sub 3 N sub 4 Up to 1000 deg C.
03 The tribological behaviour of Si sub 3 N sub 4 /Si sub 3 N sub 4 sliding pairs in pin-on-disk configuration for sliding velocities between 0.03-3 m/s, constant load of 10 N and environment-temperatures between 22-1000 deg C is dependent on the overlap ratio, the temperature and the sliding velocity. An influence of the phase composition was not observed for the three tested commercial Si sub 3 N sub 4 materials. The results are: (1) Coefficient of friction lies for solid state friction under steady state conditions between 0.5-1. (2) Wear rate increases with rising ambient temperature–especially at sliding speeds < 1 m/s. (3) The tribological behaviour for temperatures => 400 deg C is characterized by a high wear/low wear transition with increasing velocities. (4) The influence of overlap ratio on wear increases with increasing ambient temperature. A small overlap ratio is tribological disadvantageous for Si sub 3 N sub 4 sliding pairs. Si sub 3 N sub 4 /Si sub 3 N sub 4 sliding pairs do not meet for the described sliding claims without lubrication.

Figure 2: Part of a document description from the materials science domain, © Materials Information. As in the epilepsy example, the present text reproduces the ASCII text of the source. The string 'Si sub 3 N sub 4' stands for the chemical formula Si_3N_4, and '=>' for the symbol '$\geq$'.

development criteria set out at the beginning of the project [31]. In this respect, Condorcet has clearly been a two-faced research project from the start: in order to build a working application (the main objective of the project), the entire indexing process had to be conceptualized first. The outcome of the latter may be regarded as Condorcet's contribution to IR, and in the long run it may prove instrumental for the more difficult and ambitious task of information extraction as well. Although we were lucky to be able to draw on substantial experience from an earlier project with this approach [30, 26], we still had to tune the results of this earlier work to Condorcet's task, and build a working prototype in accord with the design criteria.

Condorcet's approach to document indexing by employing linguistic engineering can hardly be considered new. There are many examples of IR systems in which linguistic engineering plays a prominent role – e.g., ADRENAL [20], FERRET [24], MEDLEE [10], and AIMS [17]. Not everyone is prepared to regard these contributions as being 'significant': for instance, Harman [15] asserts that linguistic engineering still has to make its first significant contribution to improving document retrieval systems. Smeaton [27] offers a reason for this perceived inadequacy: according to him, IR and linguistic engineering are inherently different processes. IR is inexact whereas linguistic engineering is not, and only a change of approaches in both IR and linguistic engineering will lead to progress, as the current approaches only cause "the 'butting of heads', which we see at present with IR attempting to cherry-pick any appropriate techniques from NLP" ([27], p. 136).

In contrast to Harman, we think the cited works do contribute to better indexing systems. We also disagree with Smeaton's opposition between IR and linguistic engineering. In our view, an index term is abstract rather than vague; see the discussion at the end of this paper. In Condorcet, then, the linguistic engineering module is tuned to the specific needs that apply to controlled-term document indexing, causing it to differ from general-purpose linguistic engineering systems. Linguistic engineering within Condorcet is highly application-oriented; the knowledge-based approach guarantees that the linguistic engineering system is based on linguistic principles, and that therefore no *ad hoc* solutions will be applied.

4.2 System design

The design criteria underlying Condorcet are mainly concerned with costs of setting up and maintaining the system, and anticipating reuse - at least of parts of the system - for tasks similar to IR, like information extraction and text summarization. This has led to a sequential modular system, in which different kinds of knowledge are used by separate parts of the indexing system, which is depicted in figure 3. To anticipate indexing of reality-level volumes of documents, indexing and retrieving documents should be fast and robust to reduce costs of using the system to an acceptable level. Reuse of existing domain knowledge resources like UMLS is another cost-saving measure. Maintainability and extendibility are served by following the familiar principle of knowledge-based engineering to separate knowledge from the programs that use it.

The system design makes it easy to determine which knowledge contributes in what way to the overall task of indexing, and therefore the system can be optimised for the task of indexing. This will be done after evaluation of the entire sys-

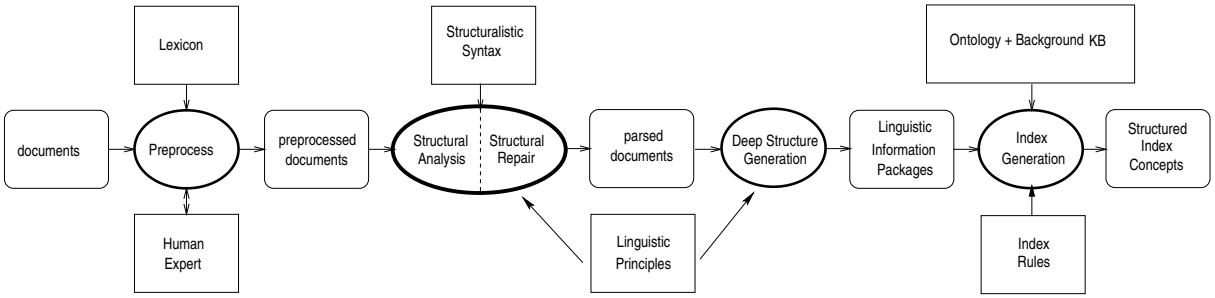

Figure 3: Condorcet's indexing process.

tem. It should be noted that this is only possible because we started out with a conceptualisation of the indexing task rather than an implementation, and because we took a rigid approach to the design of a sequential modular system, by postulating a separate module for each different type of information used. This is why the indexing process consists of several subprocesses, discussed below.

4.3 Assiging index terms

Index terms, possibly co-ordinated, are assigned to documents in a four-step process, see figure 3. The first step is pre-processing. The texts are converted from the format in which they are found into a canonical format in which SGML tags are used to delimit and identify the various parts. Further, the text is tokenised, which means that lexical units are recognised and tagged with the appropriate part-of-speech information (like plural noun, determiner, passive participle of verb, and the like). In case of lexical ambiguity, simply all possibilities are given. Tagging is based on the CELEX lexicon [4], transformed to reflect the parts of speech we want to distinguish, and on additional information for lexical items not found in CELEX. Tokenisation is interactive: when a lexical item cannot be recognised, the user is asked to supply the missing information. The pre-process is stable.

The texts are now ready for semi-automatic assignment of co-ordinated and/or unco-ordinated index terms. Obviously, the assignment will have to be based on an analysis of the natural-language text. The major problems in mapping descriptions to concepts and relations are linguistic in nature. Therefore we need knowledge on how concepts and relations can be expressed in natural language. It appears that there are many possible ways, by using different syntactic constructions. Consider the following sentences:

- Effects of zonisamide in children with epilepsy.
- Zonisamide affects epilepsy.
- Epilepsy was affected by using zonisamide.
- Zonisamide was effective in 39% of cases of epilepsy

Given the coarse granularity of the ontology used, these sentences all express the same co-ordinated index term `affects(zonisamide, epilepsy)`, only in a different syntactic form. In order to produce the proper structured conceptual representations, we not only need to determine the syntactic surface structures of these sentences but also their underlying deep structure. The deep structures contain the necessary information for mapping natural-language utterances onto terms and co-ordinators.

To obtain deep structures we use syntactic principles of Chomsky's Government & Binding (GB) theory [5]. This theory is chosen for theoretical and practical reasons. First and foremost, Chomsky's *Principles & Parameters* framework can explain a wide variety of language phenomena using just a few assumptions (see also [9]). Using GB therefore makes it possible to develop a relatively small and elegant, principle-based linguistic engineering system. As it is of secondary interest how these principles should be formalized in GB [6], we can freely formalize and implement them, and separate the linguistic knowledge resources from the processes as required.

Structural analysis

Deep structures are generated from surface structures. The Structural Analyzer produces a sur-

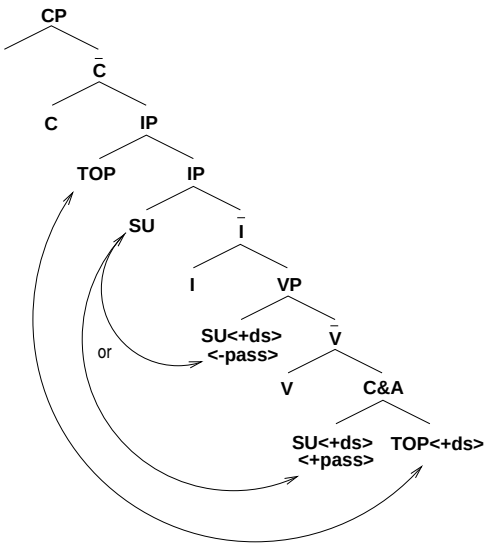

Figure 4: Enriched canonical tree structure. SU-node contains the syntactic subject and TOP collects all topicalised elements (CP's, PP's, NP's, adverbs). V contains the main verb, and C&A all verbal complements and adverbials. In enriched structures, TOP and SU are linked to their deep structure positions. SU is linked to the deep structure subject position of the main verb in case of active, and to the leftmost position under C&A in case of passive sentences.

face structure in a canonical format for every sentence in the document description, according to X-bar theory [18]. In this process, N,V, A, and P are regarded as *lexical heads*, and C(omplementizer) and I(nflection) as *non-lexical heads*. At structural level, the major categories are analysed in accord with the $\bar{X}$ Conventions. The *maximal projections* for the lexical heads are represented as NP, VP, AP, PP, IP and CP, respectively.

The Structural Analyzer, developed by Condorcet team member Erik Oltmans [25], is robust. It handles erroneous input like misspelled words and ungrammatical sentences by means of reanalysis. It has been everyday linguistic engineering practice in the last decade or so to tackle erroneous input with some robustness device, but Condorcet's Structural analyzer is unique in that it performs reanalysis in a purely principle-based fashion. It contains a number of reanalysis rules based on linguistic principles that transform partial parses (containing *chunks*) into complete parses, in accord with the canonical X-bar format. Three strategies are used in this respect: *chunking*, *sloppy agreement* and *catch-all rules* [25]. The catch-all rules ensure that the system displays behaviour known as graceful degradation. A parse is found that contains as much syntactic information as possible.

Deep structures

The next indexing step involves the generation of deep structures from the surface structures. Actually, *Enriched Surface Structures* (ESSes) rather than deep structures are generated. ESSes are constituent structures in which constituents are linked to their deep structure positions, without changing the word order of the sentence. Deep structure generation is performed by a transformational process, based on *Move* α rules and *Control Structure* rules, reflecting the principles and parameters of GB theory. A crucial condition for all *Move* α rules is that they obey the *Subjacency Condition*, thus adhering to the principle of strict cyclicity [5]. Linking constituents to deep structure positions is making use of *Case Theory* and *Theta theory*. Generating the ESS of a sentence consists of linking constituents to their deep structure positions. The result of this process is illustrated in figure 4. Deep structure positions are the deep structure subject position for the external theta role, and positions under C&A for internal theta roles.

ESS generation is complicated for a number of syntactic constructions. Consider infinitival clauses lacking overt syntactic subjects, like in the sentence *"The purpose was to inquire into the determinants of psychopathology"*. It is the task of deep structure generation to make the semantic subject explicit.

Linguistic Information Packages

After ESS's have been generated, *Linguistic Information Packages* (LIP's) can be generated for all XPs in the sentence, in a simple fashion. A LIP consists of the head of the (lexical) XP, and the heads of the lexical XPs that are in theta role positions (i.e., subject position and object position) of the matrix XP (see figure 5). LIPs contain the essential linguistic information from which structured concepts can be derived by using domain and background knowledge only. In other words, once LIPs have been generated for all XPs in a sentence, no further inferencing using linguistic knowledge has to be performed.

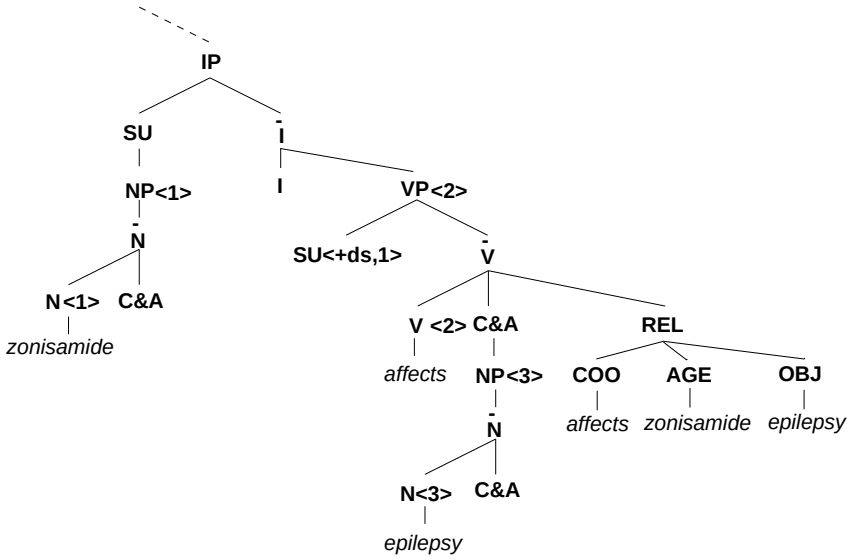

Figure 5: Syntactic tree containing the Linguistic Information Package for the VP of *zonisamide affects epilepsy*. The LIP consists of a candidate *relation* (REL), consisting of a candidate *co-ordinator* (COO), an *agent* (AGE) and an *object* (OBJ). Constituents are linked to deep structure positions by indexes (e.g. `<1>`).

Index generation

LIP's are passed on to the last module of Condorcet, called 4MOD. It consists of a knowledge base and an index generator. The knowledge base consists of four parts: (1) a lexicon that maps natural-language words (and occasionally phrases) onto thesaurus terms; (2) a list of terms; (3) a semantic network; and (4) a rule base. With the exception of the rule base, all components are taken over from the UMLS knowledge sources. For the materials science domain, there is no knowledge resource comparable to UMLS and Condorcet's knowledge base for this domain is in an embryonic stage. The rule base contains mainly disambiguation knowledge to select, for instance, the correct interpretation among the many potential interpretations of 'of': compare "effects of zonisamide" (co-ordinator `affects` with `zonisamide` as first argument and no apparent second argument) and "the inferior colliculus of rats" (co-ordinator `part-of` with `inferior colliculus` as first argument and `rat` as second argument).

The index generator takes a LIP as input. It first makes explicit all possibilities implicitly coded in the LIP. Then, if there are several possibilities, a selection process is started to rule out candidates. The knowledge is supplied by the knowledge base. For example, suppose we have received a LIP with a head that, according to the concept lexicon, might give rise to the co-ordinator `affects`. The index generator will now search the semantic subject and semantic object positions in the LIP for strings that give rise to index terms. If found, those index terms may be arguments to the co-ordinator just found, `affects`. However, the terms that are allowed as first and second argument of this co-ordinator have to be of specific types. Using the type assignments, this is checked. If the candidate co-ordinated term meets both linguistic and type-compatibility criteria, the interpretation is judged correct. Else, it is discarded and a search starts for a new interpretation.

When all LIPs have been processed, the result is polished. Duplicates are removed, as are index terms that are superterms of terms also on the list.

4.4 Provisional results

We are now in the last year of the project. Provisional testing has yielded promising results. The indexing system except the last module is able to process the larger part of the development corpus. Only the last module of Condorcet is still under development. We still need to conduct evaluation

experiments.

We find that combining linguistic and knowledge-based engineering strategies in document indexing is a viable strategy. Especially the use of substantial linguistic engineering has paid off, even though not all possible linguistic structures (adposition, extraposition, to name a few) are covered by the system. We think that coverage of these structure types is not needed for indexing purposes. We cannot substantiate this, however, because we are unable to compare the current approach with one in which these structure types are covered. We expect that for the more challenging task of information extraction, we will have to cover these structure types.

5 Discussion

Above, we have observed that one of the salient advantages of controlled terms is their independence of the language in which the document happens to be written. But the semi-automatic indexing system that assigns such terms cannot be language-independent, at least not entirely. The ideal is a system with clearly separated language-dependent and language-independent modules. Switching from English to Japanese texts would then require replacement of the modules for English by their counterparts for Japanese while the rest of the system remains unaffected.

Here the questions start. One is: is this modular design possible? Doubts are raised by observing that certain tasks need both linguistic and domain-related knowledge. In an earlier study [29], we demonstrated that anaphora resolution improves by having the program take recourse to domain knowledge in addition to linguistic knowledge. In Condorcet, disambiguation of PP-attachment is performed using UMLS constraints on relations, surely domain knowledge. Thus, it is simple to keep linguistic and domain-related resources apart but it is an open question whether the programs that use both kinds of resources can be ported to other languages without difficulty.

The combination of linguistic engineering and knowledge-based engineering is fascinating in its own right. Earlier publications (like [11, 1]) have approached the subject from the linguistic point of view. Like other work we have done in this direction, Condorcet approaches the subject from the point of view of the application to be built. The issue then becomes one of selecting resources. Sticking to the Condorcet example, for any text the search space is formed by all controlled terms (co-ordinated or not) defined by the ontology. The analysis steps are there to make constraints explicit. The constraints narrow down the search space, eventually leaving only those terms that can be assigned legitimately. (See [14] for a similar approach.) This view treats linguistic and domain knowledge as being completely on a par, without any pre-defined sequence or priority. This way of viewing the problem raises a host of interesting research questions. One of our favourites is: would it help (be more effective, be more efficient) to perform a tentative mapping on a knowledge representation first and perform linguistic analysis only later to narrow down the remaining possibilities?

Another direction in which this work can be extended is that of information extraction. From the Condorcet point of view, the difference between assigning controlled terms and transforming a text into a knowledge representation is gradual. Controlled terms, particularly co-ordinated terms, are viewed as knowledge representations that abstract from what the text actually asserts about the subject. To illustrate, `cures(aspirin, headache)` is a controlled term while ¬`cures(aspirin, headache)` ("aspirin does not help against headache") or `cures(aspirin, headache, 85, human)` ("aspirin cures headache in 85% of human patients") are ways to express what the text asserts. The Condorcet system can be enhanced to deliver the latter kind of output, turning it into an information extraction system.

Information extraction is the inevitable successor of information retrieval that, in the way we have discussed it, is better called document retrieval. A document is a combination of content and wrapper. Now information is exchanged over networks, the wrapper stands in the way of reuse of information by desktop applications of the user. We do not claim that each and every message can be couched in a knowledge representation language while preserving the nuances and modalities. But in particular in the natural sciences and engineering there is a growing need for information that is exchanged in more formal languages. The forerunners here are molecular biologists, who exchange genetic information in a form that facilitates reuse by the receiver. There is even a system that supports peer review of electronically exchanged genetics findings [8]. At the moment, however, the majority of researchers stick to articles as their mode of communication. Information extraction can be used retrospectively and concurrently to make the in-

formation available in a form fit for computer manipulation that augments the article itself. To make this possible, we need a more thorough understanding of the delicate interplay of linguistic and domain-related knowledge.

Acknowledgements

The ideas discussed in this paper have been formed in discussions with Condorcet team members Reinier Boon, Nicolaas Mars, and Erik Oltmans and former team member Jeroen Nijhuis. Other people who have contributed in one way or another, sometimes unwittingly and perhaps unwillingly, are Harold Boley, Peter Bosch, Theo Huibers, Franciska de Jong, John Mackenzie Owen, Gerrit van der Veer, and Arjen de Vries.

References

[1] James F. Allen. Natural language, knowledge representation, and logical form. In Madeleine Bates and Ralph M. Weischedel, editors, *Challenges in natural language processing*, pages 146–175. Cambridge University Press, Cambridge, 1993.

[2] David C. Blair. STAIRS redux: thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, 47:4–22, 1996.

[3] David C. Blair and M.E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28:289–299, 1985.

[4] Gavin Burnage. *CELEX - A guide for users*. Centre for Lexical Information, Nijmegen, The Netherlands, 1990.

[5] Noam Chomsky. *Lectures on government and binding*. Foris Publications, Dordrecht, The Netherlands, 1981.

[6] Noam Chomsky. On formalization and formal linguistics. *Natural Language and Linguistic Theory*, 8:143–147, 1990.

[7] William S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24:87–100, 413–424, 1973.

[8] Jérôme Euzenat. Building consensual knowledge bases: context and architecture. In Nicolaas J.I. Mars, editor, *Towards very large knowledge bases. Knowledge Building and Knowledge Sharing 1995*, pages 143–155. IOS Press, Amsterdam, 1995.

[9] Sandiway Fong. *Computational properties of principle-based grammatical theories*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass, 1991.

[10] C. Friedman, G. Hripsak, W. DuMouchel, S.B. Johnson, and P.D. Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1:83–108, 1995.

[11] Peter Gerstl. Linking linguistic and non-linguistic information. *Data and Knowledge Engineering*, 8:205–222, 1992.

[12] Thomas R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43:907–928, 1995.

[13] Thomas R. Gruber and Gregory R. Olsen. An ontology for engineering mathematics. In Jon Doyle, Erik Sandewall, and Pietro Torasso, editors, *Principles of knowledge representation and reasoning: proceedings of the fourth international conference (KR'94)*, pages 258–269, San Francisco CA, 1994. Morgan Kaufmann.

[14] Udo Hahn and Klemens Schnattinger. Ontology engineering via text understanding. In José Cuena, editor, *IT KNOWS (Information technology and knowledge systems). Proceedings of the XV. IFIP World Computer Congress, 31 August – 4 September 1998, Vienna/Austria and Budapest/Hungary*, pages 429–442, Vienna, 1998. Österreichische Computer Gesellschaft.

[15] Donna Harman, Peter Schäubele, and Alan Smeaton. Document processing. In Giovanni Battista Varile and Antonio Zampoll, editors, *Survey of the state of the art in human language technology (http://www.cse.ogi.edu/CSLU/HLTsurvey/)*. Centre for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 1996.

[16] William R. Hersh. *Information retrieval: a health care perspective*. Springer, New York, 1996.

[17] Julia Hodges, Shiyun Yie, Ray Reighart, and Lois Boggess. An automated system that assists in the generation of document indexes. *Natural Language Engineering*, 2:137–160, 1996.

[18] Ray Jackendoff. $\bar{X}$ *syntax: a study of phrase structure*. MIT Press, Cambridge, Mass, 1977.

[19] Leonardus Laeyendecker. *Brengt de vooruitgang ons verder?* Ten Have, Baarn, 1986.

[20] David D. Lewis, W. Bruce Croft, and Nehru Bhandaru. Language-oriented information retrieval. *International Journal of Intelligent Systems*, 4:285–318, 1989.

[21] D.A.B. Lindberg, B.L. Humphreys, and A.T. McCray. The unified medical language system. *Methods of Information in Medicine*, 32:281–291, 1993.

[22] Nicolaas J.I. Mars. What is an ontology? In Alex Goodall, editor, *The impact of ontologies on reuse, interoperability and distributed processing*, pages 9–19. Unicom, Uxbridge, Middlesex, UK, 1995.

[23] Sabine Martin and Günter Bergerhoff. Chemical abstracts online: a study of the quality of controlled terms. *Journal of Chemical Information and Computer Sciences*, 31:147–152, 1991.

[24] Michael J. Mauldin. *Conceptual information retrieval. A case study in adaptive partial parsing*. Kluwer Academic, Boston, 1991.

[25] Erik Oltmans. A two-stage model for robust parsing. In Chadia Moghrabi, editor, *Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA'98)*, pages 233–239, Moncton, New Brunswick, Canada, 1998. GRÉTAL, Université de Moncton.

[26] Geert J. Postma, B. van Bakel, and G. Kateman. Automatic extraction of analytical chemical information. system description, inventory of tasks and problems, and preliminary results. *Journal of Chemical Information and Computer Science*, 36:770–785, 1995.

[27] Alan F. Smeaton. Information retrieval: still butting heads with natural language processing? In M.T. Pazienza, editor, *Information Extraction - A multidisciplinary approach to an emerging information technology*, pages 115–138. Springer, Berlin, 1997.

[28] Karen Sparck Jones. Summary performance comparisons TREC-2, TREC-3, TREC-4, TREC-5, TREC-6. In Ellen M. Voorhees and Donna K. Harman, editors, *The sixth text retrieval conference (TREC-6)*, pages B–1 – B–8, Gaithersburg MD, 1998. U.S. Department of Commerce, National Instititute of Standards and Technology.

[29] Laudy E.H.M. ter Haar, Ivana Korbayová, Paul E. van der Vet, and Toine Andernach. Use of domain knowledge in resolving pronominal anaphora. *Belgian Journal of Linguistics*, 10:12–35, 1996.

[30] Bas van Bakel. *A linguistic approach to automatic information extraction*. Ph.D. thesis, University of Nijmegen, The Netherlands, 1996.

[31] Bas van Bakel, Reinier T. Boon, Nicolaas J.I. Mars, Jeroen Nijhuis, Erik Oltmans, and Paul E. van der Vet. Condorcet annual report. Technical report UT-KBS-96-12, University of Twente, Enschede, The Netherlands, September 1996.

[32] Paul E. van der Vet and Nicolaas J.I. Mars. Bottom-up construction of ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 10(4):513–526, 1998.

[33] Paul E. van der Vet and Nicolaas J.I. Mars. CQE: a query engine for coordinated index terms. *Journal of the American Society for Information Science*, forthcoming, 1999.

[34] Ellen M. Voorhees and Donna K. Harman. Overview of the sixth text retrieval conference (TREC-6). In Donna K. Harman, editor, *The sixth text retrieval conference (TREC-6)*, pages 1–24, Gaithersburg MD, 1998. U.S. Department of Commerce, National Instititute of Standards and Technology.